GLA in MediaEval 2018 Emotional Impact of Movies Task

Jennifer J. Sun[1], Ting Liu, Gautam Prasad
Google LLC
jjsun@caltech.edu,{liuti,gautamprasad}@google.com

ABSTRACT

We present our methods for the MediaEval 2018 Emotional Impact of Movies Task to predict the expected valence and arousal continuously in movies. Our approach leverages image, audio, and face based features computed using pre-trained neural networks. These features were computed over time and modeled using a gated recurrent unit (GRU) based network followed by a mixture of experts model to compute multiclass predictions. We smoothed these predictions using a Butterworth filter for our final result.

1 INTRODUCTION

The Emotional Impact of Movies Task [7], part of the MediaEval 2018 benchmark, provides participants with a common dataset for predicting the expected emotional impact from videos. We focused on the first subtask in the challenge: predicting the expected valence and arousal continuously (every second) in movies. The dataset provided by the task is the LIRIS-ACCEDE dataset [3, 4], which is annotated with self-reported valence and arousal every second from multiple annotators. Since deep neural networks, such as Inception [19], have millions of trainable parameters, the competition data may be too limited to train these networks from random initializations. Therefore, we used networks that were pre-trained on larger datasets, such as ImageNet, to extract features from the LIRIS-ACCEDE dataset. The extracted features were used to train our temporal and regression models.

This task is recurring with multiple submissions every year [11, 14]. Our method's novelty lies in the unique set of features we extracted including image, audio, and face features (capitalizing on transfer learning) along with our model setup, which comprises of a GRU combined with a mixture of experts.

2 APPROACH

We approached the valence and arousal prediction as a multivariate regression problem. Our objective is to minimize the multi-label sigmoid cross-entropy loss and this could allow the model to use potential relationships between the two dimensions for regression. We first used pre-trained networks to extract features. To model the temporal aspects of the data, the methods we evaluated included long short-term memory (LSTM), gated recurrent unit (GRU) and temporal convolutional network (TCN). Multiple modalities were fused with late fusion, and valence along with arousal was predicted jointly. Our method is implemented using TensorFlow and we used the Adam optimizer [12] in all our experiments.

2.1 Feature Extraction

We extracted image, audio and face features from each frame of the movies. Our image features (Inception-Image) were from the Inception network [19] pre-trained on ImageNet [16]. We extracted audio features using AudioSet [9], which is a VGG-inspired model pre-trained on YouTube-8M [1]. For the face features (Inception-Face), we focused on the two largest faces in each frame and used an Inception based architecture trained on faces [17]. Since the movies were human-focused, faces were found in most of the scenes. We compared these features with those used in last years competition that included image features computed using VGG16 [18] and audio features computed using openSMILE [8]. All our features were extracted at one frame per second.

2.2 Temporal Models

To model the temporal dynamics of the emotion in the videos, we used recurrent neural networks. In particular, we used LSTMs [10] and GRUs [6] as part of our modeling pipeline in a sequence-to-one setup with sequence length of 10, 30 or 60 seconds. The self-reported emotions likely depend on past scenes in movies, so temporal modeling is important for this task. In addition, we evaluated TCNs because of their promising performance in sequence modeling [2] and action segmentation [13]. Specifically, we trained an encoder-decoder TCN using sequences of extracted features to obtain a sequence of valence and arousal predictions.

2.3 Regression Models

The input from each modality (image, audio, or face) is fed into separate recurrent models. The output we use from each recurrent model is its hidden state which contains information on previous data seen by the model. We use the hidden state corresponding to the final timestamp in the input sequence. The state vectors for each modality are concatenated into a single vector and fed into a context gate [15]. Multimodal fusion occurs at this stage as we use the learned model to fuse the representation of each modality from the RNN. The output of the context gate is then fed into a mixture-of-experts model with another context gate to obtain the final emotion predictions. We use logistic regression experts with a softmax gating network.

To prevent overfitting, we regularized our models using L2 regularization, dropout and batch normalization. Finally, a low pass filter is applied on the predictions to smooth the

[1]This work completed during Jennifer's internship at Google.

MediaEval'18, 29-31 October 2018, Sophia Antipolis, France

prediction outputs. In LIRIS-ACCEDE, the measured emotion data vary smoothly in time but our regression outputs contain high frequency signals. To smooth our outputs, we tested weighted moving average filters and low-pass filters (Butterworth filter [5]) as implemented in SciPy.

3 RESULTS AND ANALYSIS

We optimized the hyperparameters of our models to have the best performance on the validation set, which consists of 13 movies from the development set. We then trained our models on the entire development set to run inference on the test set. Our setup used a batch size of 512.

Through evaluating our recurrent models, we found that Inception-Image+AudioSet features had better performance in terms of MSE and PCC compared to VGG16+openSMILE features. In some cases, the recurrent model would predict near the mean for both valence and arousal while using VGG16+openSMILE. This may be because the features did not have enough information for the models to discriminate between different values of valence and arousal. We also found a significant increase in performance when we added the Inception-Face features, which may point to salient information captured in connection with the expected emotions in the videos.

The sequence-to-one recurrent models worked best with longer input sequences of 60 seconds versus those of 10 or 30 seconds. This may be because the invoked emotion is affected by longer lasting scenes. Our recurrent models also performed better on the validation set than the TCN and the GRU models had similar performance to the LSTMs. We used GRUs for our implementation because GRUs are computationally simpler than LSTMs. Since we have a small dataset, we wanted to reduce model complexity to prevent underfitting. Our temporal model architecture ranged from 32 to 256 units and 1 to 2 layers, optimized for each of the modalities. For post processing, the low-pass Butterworth filter worked better than the moving average filter. This is likely because the Butterworth filter is designed to have a frequency response as flat as possible (with no ripples) in the pass-band. Fluctuations of the magnitude response within the passband may decrease the accuracy of our regression output.

In Table 1 we list the performance of our best models that were submitted to the task. Each of the 5 runs is defined as follows where we used Inception-Image, AudioSet, and Inception-Face as the features and a GRU with mixture-of-experts for regression.

(1) No dropout or batch normalization.
(2) Regularized with dropout and batch normalization. Trained on approximately 70% of the data.
(3) Regularized with dropout and batch normalization.
(4) Regularized with dropout and batch normalization, different initialization and epoch.
(5) Average over all runs.

We see that creating an ensemble from our models by averaging over the runs has the lowest MSE (Run 5). This is likely because by averaging, we decrease the variance of predictions and thus overall, the mean is closer to the ground truth labels. While averaging improves MSE, it does not improve correlation. Our model with the best correlation is from Run 4.

Table 1: Performance of our five models.

	Valence		Arousal	
	MSE	PCC	MSE	PCC
Run 1	0.1193	0.1175	0.1384	0.2911
Run 2	0.0945	0.1376	0.1479	0.1957
Run 3	0.1133	0.1883	0.1778	0.2773
Run 4	0.1073	**0.2779**	0.1396	**0.3513**
Run 5	**0.0837**	0.1786	**0.1334**	0.3358

We note that using batch normalization during inference increases the variance of our predictions. This is because we are using the batch statistics instead of the population statistics from the train set to normalize the batches. Our validation results (with repeated runs) as well as test results show that using batch normalization in this way improves predictions for valence, but not as much for arousal. This is most likely because the statistics of the test set for valence is different from train set while the test set statistics for arousal may be closer to the train set statistics. One explanation could be the small size of the dataset so that the statistics of the train set does not generalize well to the test set.

4 CONCLUSIONS

We found that precomputed features modeling image, audio, and face in concert with GRUs provided the optimal performance in predicting the expected valence and arousal in movies for this task. Based on our test set metrics, ensemble methods such as bagging could be useful for this task.

We found some evidence that recurrent models performed better than TCN. However since we only evaluated the encoder-decoder TCN more investigation will be necessary for a broader conclusion.

The pre-computed features we used to model image, audio, and face information showed better performance when compared with the VGG16+openSMILE baseline. A future direction could be to train the network in an end-to-end manner to better capture the frame level data, with the caveat that we may need a much larger training dataset.

REFERENCES

[1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. 2016. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675* (2016).

[2] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. 2018. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271* (2018).

[3] Yoann Baveye, Emmanuel Dellandréa, Christel Chamaret, and Liming Chen. 2015. Deep learning vs. kernel methods: Performance for emotion prediction in videos. In *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*. IEEE, 77–83.

[4] Yoann Baveye, Emmanuel Dellandrea, Christel Chamaret, and Liming Chen. 2015. Liris-accede: A video database for affective content analysis. *IEEE Transactions on Affective Computing* 6, 1 (2015), 43–55.

[5] Stephen Butterworth. 1930. On the theory of filter amplifiers. *Wireless Engineer* 7, 6 (1930), 536–541.

[6] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).

[7] Emmanuel Dellandréa, Martijn Huigsloot, Liming Chen, Yoann Baveye, and Mats Viktor Sjöberg. 2018. The mediaeval 2018 emotional impact of movies task. In *MediaEval 2018 Multimedia Benchmark Workshop Working Notes Proceedings of the MediaEval 2018 Workshop*.

[8] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. 2013. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 835–838.

[9] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 776–780.

[10] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[11] Zitong Jin, Yuqi Yao, Ye Ma, and Mingxing Xu. 2017. THUHCSI in MediaEval 2017 Emotional Impact of Movies Task. *Proc. MediaEval* (2017).

[12] Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[13] Colin Lea, Rene Vidal, Austin Reiter, and Gregory D Hager. 2016. Temporal convolutional networks: A unified approach to action segmentation. In *European Conference on Computer Vision*. Springer, 47–54.

[14] Yang Liu, Zhonglei Gu, and Tobey H Ko. 2017. HKBU at MediaEval 2017 Emotional Impact of Movies Task. In *Mediaeval 2017 Workshop. Dublin, Ireland*.

[15] Antoine Miech, Ivan Laptev, and Josef Sivic. 2017. Learnable pooling with Context Gating for video classification. *arXiv preprint arXiv:1706.06905* (2017).

[16] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, and others. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision* 115, 3 (2015), 211–252.

[17] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 815–823.

[18] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[19] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2818–2826.